Towards a Model Based Paradigm for Efficient Coding of Context Dependent Video Material

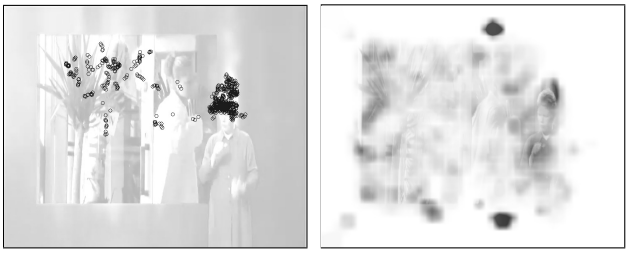
S.J.C. Davies*, D. Agrafiotis, C.N. Canagarajah, D.R. Bull

Centre for Communications Research, University of Bristol, Woodland Road, Bristol, BS8 1UB, UK

E-mail: sam.davies@bris.ac.uk

Abstract

This paper proposes a generalised framework for model based, context dependent video coding, based on exploitation of characteristics of the human visual system. The system utilises variable quality coding, based on a map which is created using context dependent rules. The technique is demonstrated for a specific video context, namely open signed content. Consequently a model for gaze prediction in open signed content is developed, based upon motion and shot changes. The framework is shown to achieve a considerable improvement in coding efficiency for the given context.



(a) Eye fixations for the sequence plotted onto one frame.

(b) One frame showing results from saliency for the entire sequence. The visible regions are those calculated to be salient across the sequence.

Figure 1. Frame 473 from the sequence 'salon'. Demonstrating the disagreement between saliency and eye tracked data.

1 Introduction

Television and online videos form an integral part of the day to day lives of large numbers of people, whether it be for entertainment, education or one of a multitude of other uses. Current standards in video coding are context independent - i.e. no high level knowledge of the content of the video is used in the coding process. This clearly has drawbacks, since applying the same coding rules to a football match and a music video cannot result in the most highly optimised coding strategy.

Previous work involving model based video coding has frequently involved the notion of saliency [6], [2]. Saliency attempts to estimate the gaze pattern for a video sequence by combining maps from a set of low-level features derived from the video (such as an intensity, colour contrast, orientation, motion etc.) into a saliency map, and then applying an algorithm which attempts to simulate the saccades of human eyes. The major problem with saliency as a gaze prediction system is that it is completely context independent, i.e. prior knowledge of the task the viewer is performing is ignored.

Saliency doesn't perform particularly well in the open signed context. Work has been done on coding sign language in a videoconference environment [5], [1], but not with open signed content. This type of video is currently utilised in broadcast scenarios, allowing sign language users to comprehend the speech, in a more natural manner than is offered by captions. The programme video is usually rescaled as an inset and the signer overlaid in the bottom right hand corner (see figure 3(a)). Figure 1 compares saliency with eye-tracked locations. The eye tracking results (fig. 1(a)) show that there is a well-defined hotspot concentrated around the face of the signer, with a few saccades over to the programme inset. The saliency plot (fig. 1(b)) doesn't agree with the eye tracking data particularly well, making it an inappropriate model for varying the quality across a frame.

We propose a framework for a generalised model based coding scheme, before utilising this framework for a specific context - namely open signed content. An initial eye tracking study will be presented, demonstrating that the context can be shown to have a gaze location pattern, before a model which will attempt to simulate this is created. This model will lead to a gaze prediction map, which will be used by a coder to generate a variable quality coded video.

This paper is organised as follows: section 2 introduces a context based coding framework. Section 3 goes on to present some eye tracking work, sections 4, 5 and 6 detail the way in which the different stages of the framework were constructed for a specific context. Section 7 demonstrates

*funded by the British Broadcasting Corporation (BBC)

Eight International Workshop on Image Analysis for Multimedia Interactive Services(WIAMIS'07)
0-7695-2818-X/07 $20.00 © 2007 IEEE

some preliminary results, before the paper is concluded in section 8, including a discussion on further work.

2 Proposed Framework

A generalised framework (see figure 2) is proposed which will overcome the problems experienced by saliency in gaze prediction. The fundamental principle of the framework is that different contexts will utilise different gaze prediction techniques. Firstly the video undergoes a context categorisation (either manual or automatic). In a broadcast scenario there is often large amounts of meta-data available alongside the video content itself which can be used to guide the context classification. Once the context is known, the framework would invoke the video pre-processing techniques applicable to that context. These could be any number of things, including saliency, ball tracking, face detection and motion estimation. A gaze prediction routine would use all of this data to estimate the most likely gaze locations for the video sequence, producing a variable quality map, which is passed to the coder. The coder uses this map to assign bits appropriately across the sequence, hence improving the perceived quality for a given bit rate, or lowering the bit rate for a given perceived quality.

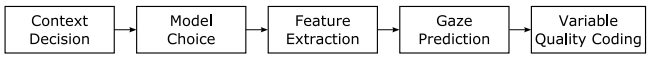


Figure 2. The proposed context driven coding framework

3 Eye Tracking Study

There was a total of eleven participants in the study, all of whom were fluent in British Sign Language (BSL). Video material was sourced from BBC open signed output, and represents current signed broadcast material. Each clip is 750 frames in length, with a frame rate of 25 frames per second. The clips were displayed at standard definition (SD) - 720x576.

Eye tracking was accomplished using an Eyelink system and in order to get just one gaze location per frame, the median of the 10 samples per frame is taken.

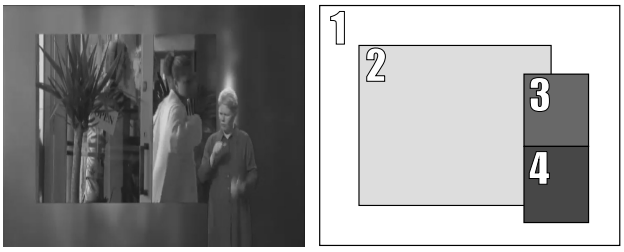
3.1 Results of Eye Tracking

Figure 1(a) shows a plot of all the eye gaze locations for the entire sequence plotted on one reference frame, for the sequence 'salon'. This clearly indicates that there are two important regions in open signed material - the signers face (as shown by Agrafiotis et al [1]), and the inset programme video. It is therefore sensible to create a gaze prediction model which will assign quality levels between these two pre-determined regions of the video.

4 Choice of Model Parameters

Observation of results suggests that gaze location is based on, in its simplest form, motion of the signer and shot changes in the inset programme video: i.e. when the signer is still, the viewers watch the inset video, and similarly when a shot change occurs in the programme, this results in a brief saccade from the signer's face to the inset.

The video sequences are segmented manually as shown in figure 3(b). In order to quantify the accuracy of any gaze prediction model a metric is generated on a frame by frame basis. This metric is simply the proportion of participants whose eye track results show they were looking at the signer for any given frame and is used to represent the Relative Region Importance (RRI) of the signer. This could be taken to represent the probability that a random sign language user would be looking at the signer for a given frame. Therefore this signal can be used to control the quality allocation across the frame.



(a) A sample frame (473) from the sequence 'salon'

(b) Segmentation regions. 1: background; 2: programme video; 3: signer's face; 4: signer's hands.

Figure 3. Region segmentation of open signed video.

4.1 Motion of Signer

The motion of the signer's hands is a clear indicator as to whether they are signing at any given time.

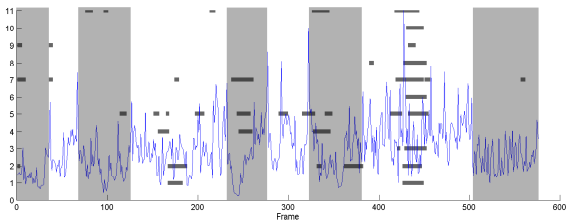
$$MS = \sum_{\{i,j\} \in R} \|mv_{i,j}\| \quad (1)$$

Equation 1 defines MS, a motion metric for the signer's hands. $mv_{i,j}$ is the h.264 [7] motion model motion vector for the (i,j)th macroblock, and R is the set of all (i,j) pairs in region 4 in figure 3(b).

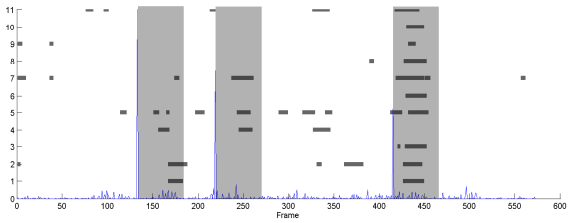
Figure 4(a) shows the motion metric and the inset fixations of each of the participants. The grey windows represent periods of lower motion. The fixations encompassed by the windows support the claim that low motion in the signer will lead to fixations on the inset.

4.2 Programme Shot Changes

To detect shot changes in the inset programme video a simple technique is employed. A 32-bin histogram generated of the intensity of the pre-determined inset region (see



(a) Motion metric (MS), with periods of low motion highlighted in light grey



(b) Shot change metric (d_2), with 50 frame windows after shot changes highlighted in light grey.

Figure 4. Horizontal dark grey bars show periods of fixation on the inset for each experiment subject. Overlaid with normalised versions of chosen metrics. Video sequence: 'salon'

figure 3(b), region 2) for each frame and the L_2 norm of the difference between consecutive frames is calculated:

$$\Delta_i = \|h_{i+1} - h_i\| = \sqrt{\sum_j (h_{i+1,j} - h_{i,j})^2} \qquad (2)$$

where h_i is the histogram vector at time i, $h_{i,j}$ is the jth bin of the histogram at time i, and Δ_i is the consecutive difference between histograms at time i.

Shot changes are most likely to be related to the 2nd difference of this motion measure. Figure 4(b) shows this normalised shot change detection metric (d_2) and the fixations on the inset for each of the test subjects. The 50 frame windows after each shot change cover fixations, which supports the idea that the shot changes cause fixations on the inset.

The majority of the fixations shown in figure 4 are 'covered' either by a shot change window, or a low motion window, implying that a combination of these metrics ought to be able to predict gaze locations.

5 Gaze Prediction Routine

$$RRI_{est}(t) = \sum_{\forall SC_i} \alpha\left(|t - t_{SC_i}|\right) + \sum_{\tau=-f/2}^{f/2} \beta_\tau MS(t-\tau) \qquad (3)$$

It is necessary to attempt to approximate the singer's RRI from the chosen metrics - shot changes in the inset video (d_2), and the motion of the signer (MS). Equation 3 defines this estimation, where SC_i is the ith shot change, at time t_{SC_i}, $\alpha(\cdot)$ is a monotonically increasing function s.t. $\alpha(x) \leq 0\ \forall x$ and β_τ are the coefficients of an $(f+1)$-tap filter.

Figure 5 shows the results of this approximation as well as the eye-tracking results the 'salon' sequence. There is a large discrepancy between sample and estimated RRI at the start of the sequence due to the fact that a viewer instinctively starts watching the signer at the beginning of any clip, and the algorithm doesn't take this into account. Here, f was taken to be 50, with β_τ defining a Gaussian filter, and the function $\alpha(\cdot)$ is as defined in equation 4 (ω, γ constants > 0). These values were chosen empirically over a series of test sequences, and then verified over a different series.

$$\alpha(x) = \frac{-\gamma}{\omega} e^{-\frac{x^2}{2\omega^2}} \quad x \geq 0 \qquad (4)$$

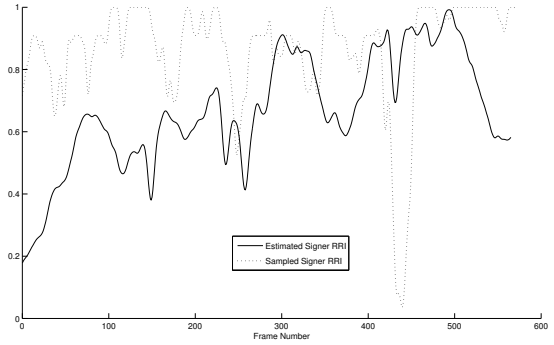


Figure 5. Estimated RRI of the signer together with sampled proportion (RRI), varying in time.

6 Variable Quality Video Coding

Having created an approximation for the signer's RRI for each frame, a variable quality map must be created. As was shown in [1] when sign language users are watching sign language, they concentrate primarily on the face and mouth, and the eye tracking data presented here support this. Therefore, if it is likely that a viewer is looking at the signer, the quality of the face should be high, and the quality of the inset programme video can be lowered. The opposite is also true.

The chosen mechanism for creating variable quality coding is to code the video using the h.264 reference coder, using the main profile, without B-Frames, and varying the quality using a variable QP map (higher QP leads to lower quality).

The QP map is comprised of 4 distinct regions demonstrated in figure 3(b) the background, the inset programme,

0-7695-2818-X/07 $20.00

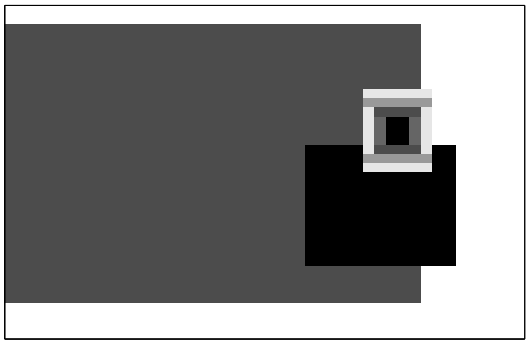


Figure 6. Sample QP map output for a frame of 'carib3'. Darker shades represent lower QP - i.e. higher quality.

the signer's face and hands. The process of assigning quality is therefore executed in 4 phases - one for each of the regions. The map is expressed as a difference map in relation to a predefined QP value. The background QP is set higher than the base level and the lower part of the signer is left the same as the base level. The programme inset and the head of the signer have QP values which depend on the model predicted RRI. The QP of the inset is constant across a frame, and is increased from the base value to a specified maximum linearly with the predicted RRI. The QP for the face of the signer is radial (to prevent visible boundaries appearing) - with the centre being low and the outside being high. A constant factor is added across the region, which is proportional to the signer's RRI estimate. Figure 6 shows a sample generated QP map.

7 Results

A subjective study was carried out, with 5 participants fluent in BSL. They were each shown a random sequence of open signed clips (standard definition) ranging in length from 10s to 40s, rating the quality of each clip out of 10 immediately after it had been shown (Single Stimulus Numerical Categorical Scale [3]). There was a total of 60 clips, half of which were coded with uniform QP, and half which followed the proposed algorithm.

$$\pi_{ij} = \frac{p_{ij} - \mu_i}{\sigma_i} \tag{5}$$

The scores were normalised as shown in equation 5, where p_{ij} is the score given by the ith participant for the jth clip, π_{ij} is the associated normalised score, and μ_i and σ_i the mean and standard deviation of the scores given by the ith participant respectively.

Figure 7 shows the difference in normalised subjective score (positive implies the proposed technique performs better) against the associated bit rate saving. It can be seen that bit rate savings of around 30% are achieveable with either no or negligible loss of perceived quality.

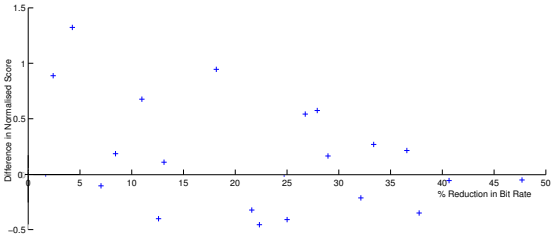


Figure 7. Subjective results: Bitrate gain vs. difference between scores for uniform and variable quality coding

8 Conclusions

A context based video coding paradigm has obvious advantages over both currently used and perceptually optimised techniques. It has been demonstrated here that there are specific contexts for which relatively simple gaze prediction routines can yield large bit-rate savings: other well-defined contexts include surveillance, sports and news.

Future improvements to this framework include work on each of the constituent parts: the context decision process could become automated, which in turn implies more models for contexts have to be developed. The feature extraction and gaze prediction select tools from a toolset depending on the chosen model - the tools used here are a small subset of those available. It would be possible to improve the variable quality coding by incorporating the gaze prediction maps into a rate control mechanism. The framework presented here is very much in its infancy, but offers huge potential for video coding.

References

[1] D. Argrafiotis, C. Canagarajah, D. Bull, M. Dye, H. Twyford, J. Kyle, and J. C. How. Optimised sign language video coding based on eye-tracking analysis. *Visual Communications and Image Processing, Proc. of SPIE*, 5150:1244–1252, 2003.

[2] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. on Image Processing*, 13(10):1304–1318, Oct 2004.

[3] ITU. Methodology for the subjective assessment of the quality of television pictures. *ITU-R BT.500-11*, 2002.

[4] O. L. Meur, D. Thoreau, P. L. Callet, and D. Barba. A spatio-temporal model of the selective human visual attention. In *ICIP*, 2005.

[5] L. Muir, I. Richardson, and S. Leaper. Gaze tracking and its application to video coding for sign language. In *Picture Coding Symposium*, Saint Malo, France, April 2003.

[6] C. Privitera and L. Stark. Algorithms for defining visual region-of- interest: Comparison with eye fixations. Technical Report UCB/ERL M97/72, EECS Dept., Univ. of California, 1997.

[7] T. Wiegand, G. Sullivan, G. Bjøntegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Trans. on Circuits and Systems for Video Technology*, 13(7):560–576, Jul. 2003.

0-7695-2818-X/07 $20.00 © 2007 IEEE